The Usage of Clustering in Hybrid Models

Christie Maria
Professor, Department of Bioinformatics, Ghent University, St. Pietersnieuwstraat 33, 9000 Gent, Belgium

Corresponding author: Christie Maria, Professor, Department of Bioinformatics, Ghent University, St. Pietersnieuwstraat 33, 9000 Gent, Belgium, Email: mariachristie81@hotmail.com

Received Date: 2nd March 2018
Accepted Date: 4th May 2018
Published Date: 5th May 2018

Researchers have proposed diverse representation models which use corpus statistics to assess the semantic similarity between documents. Some of these models focus on correlation between the terms to clearly approximate the semantic similarity while others seek to get the latent representation of concepts through application of dimension reduction techniques [1]. The researchers combined latent and explicit analysis to develop new hybrid models that can approximate the semantic similarity that exists between the documents [2]. In addition, they have been utilized to improve how the document clustering algorithms usually perform. The analysis done on the thirteen sets of benchmark data proves there are meaningful perfections in clustering when the hybrid models are used together with clustering algorithms with high sensitivity to errors in approximating document similarity [3]. Document clustering seeks to organize documents into clusters based on their topics [4]. Document clustering algorithms have been utilized since the era when text retrieval systems were popularly relied in organizing documents to enhance the efficiency and effectiveness of the retrieval system as well as facilitating end-user browsing to organize documents [5]. Currently, document clustering has gained prominence given the massive volume of documents available in open source platforms such as the Internet [6]. Researchers are interested to explore ways to enhance the success and proficiency of document clustering algorithms while they sustain scalability [7]. Different types of clustering algorithms have been used to document sets of data [8]. They include the most recent one known as non-negative matrix factorization as well as spectral clustering, k-means, and hierarchical methods [9]. Vector Space Model (VSM) is broadly applied in such algorithms exclusively as their fundamental model for document representation [10]. They evaluate the proximity between documents on the basis of their vector’s inner-product as well as space of terms [11]. Document clustering algorithms techniques such as spectral clustering and hierarchical methods can be applied to the cosine similarities’ matrix; alternatively, the non-negative matrix factorization and the k-means are usually applied straight to the document vectors [12]. Performing document clustering using the Vector Space Model as the fundamental model for document representation entirely disregards the term’s semantic relations [13]. Therefore, the documents which lack the common terms are usually regarded as being dissimilar despite containing several terms that have semantic relations [14]. Alternatively, documents with several common terms are regarded as being the same despite such common terms being noisy while other terms in the document might be semantically unrelated [15]. The need to do away with the VSM’s limitations led to the development of diverse models of document representation which capture the similarity between documents by evaluating the term occurrence patterns using statistical analysis [16]. The generalized vector space model (GVSM) is a good example of such models as it explores terms to explicitly compute measures of correlation that is evident between terms [17]. Subsequently, it utilizes the measures to approximate the similarities between documents [18-19]. Other alternative models like the latent semantic indexing (LSI) usually find the latent representation of concepts through the use of dimension reduction techniques [20]. The proposed hybrid model performs mapping of documents to a semantic space whereby similarity between documents signify extent to which their terms are statistically correlated and subsequently find an abridged representation that with the ability of preserving semantic similarity between documents by applying the dimension reduction techniques [21].

The researchers then conduct an empirical assessment between the four recognized models and the proposed models for semantic analysis then they investigate how such models improve the effectiveness of the diverse document clustering algorithms [22]. Finally, the study results indicates hybrid models are statistically equivalent or significantly better in comparison to other representation models used in capturing how documents are semantically related [23]. Some clustering algorithms techniques such as the hierarchical algorithms which depends on making local decisions is usually more sensitive to errors in approximating document similarity and consequently leverage on the hybrid models [24]. The hierarchical algorithms’ outputs is more insightful for document data sets since it certainly signify the hierarchy of topics [25]. The researchers propose that future work related to enhancement of document clustering through the use of semantic analysis must involve investigating the problem of establishing the latent space’s inherent dimensionality as well as designing new techniques that rely on using the recommended models with clustering algorithms that minimize the computational complexity related to dimension reduction and semantic mapping as well as require the non-negative matrices [26].
Table 1: Average of relative quality measures for different representation models and clustering algorithms (Source: Inderjit SD, Dharmendra SM (2001) Concept Decompositions for Large Sparse Text Data Using Clustering, Machine Learning 42: 143-175).

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Spherical k-means</td>
<td>0.9186</td>
<td>0.9668</td>
<td>0.9750</td>
<td>0.9640</td>
<td>0.9780</td>
<td>0.9676</td>
</tr>
<tr>
<td>HAC-Complete</td>
<td>0.6633</td>
<td>0.9113</td>
<td>0.8503</td>
<td>0.8943</td>
<td>0.9828</td>
<td>0.9742</td>
</tr>
<tr>
<td>HAC-Average</td>
<td>0.7109</td>
<td>0.8605</td>
<td>0.9160</td>
<td>0.9712</td>
<td>0.9662</td>
<td>0.9557</td>
</tr>
<tr>
<td>Spherical k-means</td>
<td>0.9000</td>
<td>0.9720</td>
<td>0.9706</td>
<td>0.9619</td>
<td>0.9529</td>
<td>0.9383</td>
</tr>
<tr>
<td>HAC-Complete</td>
<td>0.5941</td>
<td>0.8897</td>
<td>0.8187</td>
<td>0.8879</td>
<td>0.9732</td>
<td>0.9616</td>
</tr>
<tr>
<td>HAC-Average</td>
<td>0.6066</td>
<td>0.7887</td>
<td>0.9002</td>
<td>0.9824</td>
<td>0.9462</td>
<td>0.9493</td>
</tr>
</tbody>
</table>

References


Submit your manuscript at http://enlivenarchive.org/submit-manuscript.php

New initiative of Enliven Archive
Apart from providing HTML, PDF versions; we also provide video version and deposit the videos in about 15 freely accessible social network sites that promote videos which in turn will aid in rapid circulation of articles published with us.